Identifying the Components[*]

Matthijs van Leeuwen, Jilles Vreeken, and Arno Siebes

Department of Computer Science
Universiteit Utrecht
{mleeuwen,jillesv,arno}@cs.uu.nl

Most, if not all, databases are mixtures of samples from different distributions. In many cases, however, nothing is known about the source components of these mixtures. Therefore, many methods that induce models regard a database as sampled from a single data distribution. Models that do take into account that databases actually are sampled from mixtures of distributions are often superior to those that do not, independent of whether this is modelled explicitly or implicitly.

Transaction databases are no different with regard to data distribution. For the prototypical example, supermarket basket analysis, one also expects a mixture of different buying behaviours. Households of retired people buy different collections of items than households with young children, although overlap may exist. By extracting both the groups of people and their corresponding buying patterns, a company can learn a lot about its customers.

But, what does "different buying behaviour" mean? It certainly does not mean that the different groups should buy completely different sets of items. Also, it does not mean that these groups cannot have common frequent item sets. Rather, it means that the *characteristics* of the sampled distributions are different. This may seem like a play of words, but it is not. Sampled distributions of transaction data can be characterised precisely through the use of a pattern-based compressor.

We introduce two MDL-based algorithms that follow orthogonal approaches to identify the components in a transaction database. The first follows a model-based approach, while the second is data-driven. Both are parameter-free: the number of components and the components themselves are chosen such that the combined complexity of data and model is minimised. Further, neither prior knowledge on the distributions nor a distance metric on the data is required.

Experiments show that highly characteristic components are identified. Both algorithms are evaluated on basis of total compressed sizes and component purity, but we also look at (dis)similarities between the components and their characteristic patterns. The results show that both our orthogonal methods identify the components of the database. Visual inspection confirms that characteristic decompositions are identified.

Reference

[1] van Leeuwen, M., Vreeken, J., Siebes, A.: Identifying the Components. Data Mining and Knowledge Discovery (2009) DOI: 10.1007/s10618-009-0137-2

[*] This is an extended abstract of an article published in the Data Mining and Knowledge Discovery Journal [1].

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, p. 32, 2009.